BASES DE DATOS Y REPRESENTACIÓN DEL CONOCIMIENTO

Ernesto GARCÍA CAMARERO

ABSTRACT

The computer has three fundamental analogies with the human being (senses, memory and intelligence), but the coming out of the *data bases* announced a new form of language: the *computer language*. The data bases not only set several new technical and social problems, but moreover are modifying the traditional form of *social memory*, supported by paper, by changing it in a memory based on electronic means. This situation creates new forms of representation of knowledge to which the author gave attention in previous contributions with his SENECA project, in which computer language tries to give a synthesis of oral and written languages, with their respective advantages.

1. Introducción

La informática ha obligado a que la ciencia tenga entre sus objetos de estudio uno que, sin ser nuevo, había pasado desapercibido aunque usado: la información. De hecho, esta afirmación no es del todo correcta, si consideramos a la lingüística como el antecedente natural de la informática, y que la informática teórica no es más que la incipiente matemática del lenguaje.

En efecto, hasta la aparición de los ordenadores, la información estaba depositada en las lenguas habladas por el hombre, y la lingüística ha tratado de dar respuestas a los fenómenos del lenguaje, entendido éste como uno de los atributos que distinguen al hombre de los otros seres de la Tierra.

La aparición de los ordenadores ha hecho que la información circule, se elabore y se interprete sin necesidad de la intervención humana directa; ha hecho que surjan lenguajes artificiales y ha hecho también que se preste a la lingüística una nueva forma de abordarla usando la informática como metáfora.

Dejando de lado, por pueriles y deformadoras, las ideas antropomórficas con que a veces se ha considerado a los ordenadores, no cabe duda que su estructura tiene algunas analogías funcionales con la del ser huma-

THEORIA - Segunda Época,
Año I (1985), N° 1, pp. 293-303.

no, que podemos agrupar en tres tipos: sentidos, memoria, inteligencia. Los sentidos son los órganos del ordenador por los que percibe información, ya sea en forma simbólica (como las lectoras de tarjetas, los teclados, etc...) o como magnitudes físicas (sensores de temperatura, presión, etc...), en los que la simbolización se hace automáticamente. La memoria permite el almacenamiento (temporal o permanente) de información estructurada para facilitar su búsqueda o para permitir asociaciones o inferencias. La inteligencia (llamada artificial) está compuesta en un ordenador por los dispositivos físicos adecuados (unidades de control, aritmética, etc...) y por los programas mediante los cuales pueda elaborarse la información "percibida" o "recordada" por el ordenador. Además, el desarrollo de la robótica agrega a esta analogía músculos mediante los que ejecutar acciones físicas como respuesta a la información elaborada.

Esta nueva situación ha hecho que la psicología y la lingüística dirijan su mirada al ordenador y se interpreten dentro de la metáfora informática, se consideren como originadas por procesos de conocimiento y surja así lo que ya se llama ciencia cognitiva.

En este marco debe situarse la aparición y desarrollo de las bases de datos y considerarse éstas como el germen de una nueva forma de representación del conocimiento, que conducirá, sin duda, a la aparición de una nueva forma de lenguaje que signifique un cambio tan trascendental como lo fue el paso del lenguaje oral al lenguaje escrito, que marcó el comienzo de la Historia.

2. Bases de datos

En los sistemas informáticos se han distinguido siempre dos tipos de información: programas y datos. Los programas se construyen para resolver problemas concretos que hayan sido analizados previamente en su enunciación general y determinado su método de solución; de hecho un programa es la descripción (en un lenguaje comprensible por la máquina) del procedimiento o algoritmo mediante el que obtenemos la solución de un problema. En general los programas se escriben de forma paramétrica para que puedan aplicarse a la obtención de las soluciones de toda una familia (eventualmente infinita) de problemas, de manera que, al fijar los parámetros, el problema quede determinado y la solución correspondiente

se obtenga de forma unívoca. La colección de los valores particulares de los parámetros mediante los que determinamos un problema concreto recibe el nombre de datos del problema.

Al comienzo de la informática se prestaba casi toda la atención a la forma de los programas y a los lenguajes de programación, debido principalmente a la complejidad de los algoritmos en relación con la sencillez de los conjuntos de datos sobre los que actuaba. La organización de los datos se establecía en el interior del programa que los utilizaría.

Pero cuando un mismo conjunto de datos debía ser usado por varios programas distintos, aquél debía organizarse de forma que los datos en él contenidos, pudieran ser utilizados por los distintos programas, lo que obligaba a que los programas se construyesen teniendo en cuenta la organización de los datos. De esta forma, se hizo patente la necesidad de organizar los datos, con independencia del programa que los fuera a utilizar y así surgieron los bancos de datos.

En un principio, los criterios de organización de los datos se apoyaban en la forma de almacenamiento, de manera que los programas pudieran encontrar y recuperar los datos necesarios para su ejecución. Esta supeditación de la organización de los datos a la estructura física de la memoria del ordenador o del soporte en el que fuera a realizarse su almacenamiento, dejó patente una rigidez excesiva en su utilización y puso de manifiesto que la estructuración de los datos no debía estar determinada por la organización física de la memoria, por sus formas de acceso, ni por la naturaleza formal de los mismos, sino que debía realizarse teniendo muy en cuenta el contenido semántico de los datos para lograr una eficiente estructuración de los mismos. Así aparecen las bases de datos.

Hemos de hacer notar que con la expresión *"bases de datos"* se denomina tanto a los *sistemas de gestión de bases de datos*, como a los datos propiamente dichos, cosa que con frecuencia lleva a confusión. No significa lo mismo *base de datos* en las siguientes frases: la base de datos Mistral, la base de datos del IBH. En el primer caso se refiere a un sistema de gestión de bases de datos, es decir, a un conjunto de programas de ordenador mediante los que podemos crear, actualizar, corregir y consultar un conjunto de datos. En el segundo caso, nos referimos al conjunto de datos documentales del Instituto Bibliográfico Hispánico, representados sobre el soporte informático y que estará controlado por un

sistema de gestión de bases de datos al que no se alude. En cualquier caso, hecha esta distinción resultan inseparables ambos conceptos, ya que sería inútil tener un sistema de gestión de base de datos sin datos que gestionar, así como también es hoy impensable tener cierta cantidad de información sin un sistema informático que la organice. Así se justifica el uso de la expresión *base de datos* de forma genérica, y se deja al contexto resolver la ambigüedad en caso de que ésta se produzca.

Los sistemas de gestión de bases de datos se pueden clasificar de varias maneras. Si consideramos la estructura de los datos, podemos hablar de bases de datos jerárquicos, si su estructura es arborescente; bases en red, si se permite la conexión entre nodos de ramas distintas; bases relacionales, si utiliza el cálculo relacional como modelo de representación, etc.; si consideramos el tipo de aplicaciones podemos hablar de bases de datos genéricas y específicas; si atendemos al tipo de datos pueden ser numéricas, factuales, textuales, bibliográficas, etc.

Pero una vez construído o disponible un sistema de gestión de bases de datos, son dos las tareas a realizar: por una parte, la *creación* de la base de datos y por otra la *distribución* y explotación de la misma. La creación consiste en formar y mantener un caudal de información lista para su consulta. La distribución consiste en localizar y hacer llegar a cada interesado la información pertinente.

Por el momento, la tarea más delicada y más cara es la creación de bases de datos. *Delicada*, porque ha de diseñarse con sumo cuidado para que su organización permita una búsqueda y elaboración eficaces; porque han de definirse normas que faciliten la integración de datos generados en distintos puntos y en distintas épocas; porque el mantenimiento debe asegurar su puesta a punto para evitar obsolescencia. *Cara*, porque en la actualidad se requiera todavía una participación personal muy alta en las tareas de creación y mantenimiento.

La distribución puede realizarse por medios convencionales (correos, ...), pero la forma natural es apoyarse en las redes de información. Estas implican, por una parte, un soporte de telecomunicaciones y por otra, un soporte de programas. En Europa la red más extendida es la denominada EURONET y el correspondiente sistema llamado DIANE; en Estados Unidos son varios los sistemas de distribución, pero por citar alguno mencionaremos TELENET como red de comunicación y DIALOG como uno de

los sistemas de distribución más difundido en ambos continentes. En España existe una red de transmisión de datos denominada IBERPAC, y entre los varios sistemas de distribución citemos los PIC (Puntos de Información Cultural) gestionados por el Ministerio de Cultura.

Con estas sucintas referencias sólo queremos subrayar el hecho, bien sabido, de que las bases y redes de datos son hoy una realidad, pero indicar también que, pese a sus grandes dimensiones en algunos casos, son una realidad incipiente, en cuanto forma y dimensión, y que distan mucho de lo que serán las bases de datos en un futuro no lejano, ya que muchos son los problemas que en la actualidad tienen planteados las bases de datos. Por una parte, problemas técnicos: desarrollo de mejores soportes de información (por ejemplo el videodisco), superación de las incompatibilidades de comunicación (mejorar los protocolos), encontrar formas de representación del conocimiento más adecuadas que el lenguaje escrito. Por otra parte, problemas sociales.

Entre los problemas sociales citemos el flujo transfronteras de la información (en un momento en el que la información es la principal riqueza), el peligro que las bases de datos acarrean sobre la privacidad y sobre la manipulación de la información que reciben los ciudadanos (lo que suscita la imagen del Big Brother controlador), la delincuencia electrónica, etc. También debemos incluir aquí la expectativa de modificación de la organización urbana y de la estructura administrativa; los síntomas y necesidad de un profundo cambio del sistema educativo, ya que dejará de ser transmisión de información para convertirse en el aprendizaje de la localización de la información y en el desarrollo de la inteligencia; la alteración de los medios informativos de masas (prensa, radio, TV,..) ya que las bases de datos facilitan la existencia de redes de comunicación simétricas (o más simétricas que las actuales caracterizadas por un solo emisor y millones de receptores) y por tanto, la posibilidad de que cada usuario confeccione diariamente su periódico o su programa de TV, usando las técnicas de perfiles sobre bases de datos de noticias de prensa, de programas de TV, de películas, etc.

Hemos visto pues cómo aparecieron las bases de datos, su incipiente estado actual de desarrollo, y algunas expectativas que técnicamente son ya posibles, y económicamente lo serán en un futuro muy próximo.

Pero queremos señalar también la trascendencia que tiene el con-

siderar a las bases de datos como modelos del mundo exterior realizados de acuerdo con ciertas formas de representación del conocimiento, el gran paso que puede derivarse para la construcción de nuevos tipos de lenguajes, y el gran riesgo de confundir el modelo con la realidad.

3. Representación del conocimiento

En la actualidad la representación del conocimiento de la Humanidad se realiza a través de su memoria social constituída por la información recogida en las bibliotecas, cartotecas, hemerotecas, archivos, registros civiles, de patentes, de la propiedad, notariales, catastros, etc.; el conocimiento de cada individuo se ha construído con su propia experiencia y con aquella pequeña parte de la información social a la que ha tenido acceso.

Las actuales formas de representación del conocimiento social se basan en el lenguaje escrito, con algunos arreglos y variantes obtenidos por fórmulas, tablas, diagramas, mapas, etc. El acceso a esa información la realiza el individuo usando medios arcaicos y muy laboriosamente. El sistema educativo actual pretende facilitar el acceso a un núcleo de información básica, lográndolo sólo parcialmente y con un esfuerzo y costo social realmente alto.

Estamos asistiendo a la sustitución de la antigua memoria social soportada sobre papel, por otra soportada en medios electrónicos. Pero no es un simple cambio de soporte, aunque en ésta, su primera, fase se procede de manera mimética, como hacía Leonardo con sus ingenios para volar. Pero empieza a percibirse con claridad que están produciéndose cambios cualitativos en la forma de representar el conocimiento social.

Ya Diderot, con su Enciclopedia, pretendía conseguir que en una sola colección de libros estuviera representado todo el conocimiento, evitando redundancias innecesarias, homogeneizando léxicos y notaciones, y organizando la información para facilitar su búsqueda. En este planteo existen tres grandes hallazgos: constatar la enorme superredundancia de información, detectar la necesidad de contar con un lenguaje homogéneo y apropiado, y ver la conveniencia de disponer de sistemas de localización rápida de la información deseada.

En la actualidad se habla de enciclopedias sobre soporte electróni-

co. El uso de esta tecnología, aporta algunas ventajas operativas sobre las antiguas formas de impresión, como son la facilidad de actualización de la enciclopedia, la posibilidad de editar, a bajo costo, diccionarios enciclopédicos sobre temas correspondientes a dominios específicos, el poder usar definiciones de mayor o menor extensión según las necesidades, etc. Pero aún así sólo se aprovechan mínimamente las posibilidades que brinda la informática, ya que aún no se ha logrado superar al papel como soporte final, ni se ha utilizado toda la capacidad dinámica del ordenador que posibilita no sólo describir procedimientos sino realizarlos. Tampoco se usan formas de acceso muy diferentes de los clásicos índices impresos, aunque ahora se realicen automáticamente y a mayor velocidad.

Otros tipos de diccionarios electrónicos se están perfilando, en los que se utilizará toda la capacidad dinámica de los ordenadores. No consistirán en grandes catálogos de definiciones, más o menos minuciosas, de términos aislados, sino que su finalidad será la de representar el conocimiento como un todo. La definición de los términos particulares se realizará en función de la representación global del conocimiento, como un elemento de ese conocimiento, en lugar de la práctica actual de considerar al conocimiento global como un conglomerado, malamente estructurado, de sus partes. Las partes sólo toman significado en función de la totalidad, aunque ésta se va conformando y modificando por agregación de nuevas partes.

Para ello se debe considerar que el conocimiento es siempre algo parcial e incompleto; que debe considerarse a la contradicción como un ingrediente potenciador del conocimiento; que la incompletitud y la contradicción obliga al conocimiento a ser algo relativo y dinámico o cambiante; que toda representación del conocimiento debe fundamentarse sobre algunos elementos primitivos (no necesariamente los mismos para todos) cuya definición o asignación de significado se haga con la punta de los dedos diciendo "a eso me refiero".

Todo esto nos conduce a la necesidad de desarrollar sistemas de representación del conocimiento con más capacidad para tal finalidad de la que posee el lenguaje escrito que no ha logrado desprenderse de la estructura que tiene en común con el lenguaje hablado; por otra parte, el lenguaje escrito presenta las siguientes características poco apropiadas para la representación del conocimiento:

- gran dificultad de modificar y sintetizar la información almacenada mediante el lenguaje escrito, así como inferir nueva información a partir de la dada.

- pasividad de la información registrada mediante lenguaje escrito.

- dificultad de localizar información específica en un corpus amplio.

Para atenuar la última dificultad señalada, se utilizan en el lenguaje escrito formas no empleadas en el oral, como son los índices, sumarios, paginación, así como la descomposición en tomos, capítulos, parágrafos, con los distintivos y títulos de cada uno de ellos para permitir su identificación; y a macroescala surge toda la actividad bibliotecaria, bibliográfica y documental.

La pasividad de la información escrita la entendemos en el sentido de que sólo ésta adquiere significado si es interpretada por el hombre, de que sólo es utilizable si es trasvasada al cerebro humano y éste realiza acciones acordes con ella.

Por otro lado, el lenguaje escrito no permite modificar con facilidad los contenidos expresados mediante su uso, ni sintetizar dichos contenidos y mucho menos obtener información implícitamente, es decir, obtener información subyacente e inferible a partir de la dada explícitamente.

La informática está determinando en la actualidad una nueva forma de representación del conocimiento (uno de cuyos ensayos hemos desarrollado y denominado SENECA). En algún sentido estamos presenciando la sustitución del lenguaje escrito por el lenguaje informático. No debe entenderse aquí lenguaje informático en el sentido restringido de los "lenguajes de programación", ni siquiera en el algo más amplio de "lenguaje utilizado por los informáticos", sino considerarlo como un sistema de representación de conocimiento usando ordenadores. Un sistema que tiene estructuras similares a las de los lenguajes oral y escrito, no sólo en su finalidad sino también en su semántica; de alguna manera representa una síntesis de los lenguajes oral y escrito. De éste tiene la característica de su perdurabilidad y de su transmisividad a distancia; de aquel (el oral) tiene su aspecto dinámico; además contiene elementos que son más propios de la inteligencia humana que de su lenguaje.

Las definiciones dadas en los diccionarios actuales adolecen del sesgo producido por los redactores de las mismas, no consiguiéndose nunca

(pese al hecho atenuante de introducir diversas acepciones) definiciones que se ajusten a situaciones concretas en las que sólo nos interesan ciertos rasgos semánticos, o bien no se logran definiciones con el grado de precisión y detalles deseado, que es distinto cada vez que se consulta el diccionario, según el uso que se ha de hacer del término; y mucho menos se logran diccionarios que describan situaciones de cierta complejidad variable según los casos.

Toda esta situación nos hace ver como evidente que la forma actual de producción escrita ha de sufrir profundas modificaciones y un nuevo lenguaje se está gestando: el lenguaje informático. Porque es obvio que utilizar la informática para transcribir en sus dispositivos mensajes codificados del lenguaje escrito es un anacronismo mediante el que se desperdicia toda la potencialidad dinámica de la nueva tecnología. Así, es de esperar que no dure mucho tiempo esa forma de difusión de la ciencia que es la prensa escrita, que produce anualmente millones de páginas escritas de un contenido altamente redundante y tiene, además, la propiedad de dificultar enormemente la localización de lo poco nuevo que hay en ellas. Todo esto produce ese gran esfuerzo de localizar la información impresa que nos interese sobre un tema particular y de ésta extraer las ideas que realmente precisamos. Por eso creemos que no se trata de crear grandes sistemas de documentación, sino más bien racionalizar la producción escrita para evitar que aquéllos sean necesarios.

Todo esto implica la *aparición* de un nuevo lenguaje, que sustituya al actual lenguaje escrito, que facilite la representación del conocimiento, y su adquisición y su uso; que evidentemente habrá que aprender, aunque esperemos que con menos costo y esfuerzo que el empleado en la actualidad para aprender a leer y a escribir.

REFERENCIAS

ANDERSON, J.R.: Induction of augmented transition networks, "Cognitive Science", 1, 1977, pp. 125-157.

BLOC, L. (Ed.): Natural Language Communication with Computers, New York, Springer, 1978.

BOBROW, D.G. e COLLINS, A.M. (Eds.): Representation and Understanding: Studies in Cognitive Science, New York, Academic Press

1975.

BOBROW, R.J. e WEBBER, B.L.: Knowledge Representation for syntactic/semantic Processing, 1st AAAI, 1980, 316-323.

BORKIN, S.A.: Data models: a semantic approach for data base systems, MIT Press, 1980.

CODASYL PROGRAMMING LANGUAGE COMMITTEE, Database Task Group Report, ACM, 1971. (Hay una versión más reciente de 1978).

CODD, E.F.: Extending the database relational model to capture more meaning, "ACM Transaction on Database Systems", vol. 4, nº 4, 1979.

COLMERAUER, A.: Metamorphosis Grammar. Natural Language Communication with Computers, Berlín, Springer, 1978.

CHAMBERLIN, D.D. e BOYCE, R.F.: Relational database management: a survey, "ACM Computing Survey", vol. 8, nº 1, 1976.

CHAMBERLIN, D.D. e BOYCE, R.F.: SEQUEL: A structured English Query Language, Proc. 1974 ACM SIOFIDET Workshop on Data Description, Access and Control.

DATE, C.I.: An introduction to data base systems, Addison-Wesley, 1981.

EPSTEIN, M.N.: Natural Language Access Clinical Data Bases, Ph. D. Th. Medical Information Sciences, University of California, San Francisco, 1980.

GARCIA CAMARERO, E., VERDEJO, M.F. e GARCIA, J.: SENECA: SEmantic networks for conceptual analysis. Data Bases in the Humanities and Social Sciences, Amsterdam, North-Holland, 1980, pp. 67-71.

GARCIA CAMARERO, E.: SENECA: un método de representación del conocimiento. Représentation des Connaissances et raisonnement dans les sciences de l'homme, Paris, INRIA, 1980, pp. 223-237

GARCIA CAMARERO, E.: Lenguaje y representación del conocimiento. Ordenadores y Lengua Española (1981), Pisa, Giordini Editori, pp. 113-120.

HENDRIX, G.G. e WILLIAM, L.: Transportable Natural Language Interfaces to Data bases, Proc. 19th. Annual Meeting of the ACL, SRI International, Menlo Park, California, 1981.

JONES, M.A.: Toward and Induction System for Conceptual Representations Foundational Inreads in Natural Language, Processing Workshop, SUNY. Stony Brook, 1981.

LESMO, L. e TORASSO, P.: Knowledge Source Co-Operation in a Natural Language Query System, ISI Internal Report, Universitâ

di Torino, Oct. 1981.

SALVETER, S.: Inferring Conceptual Graphs, "Cognitive Science", 3, 2, 1979.

SCHANK, R.C. e COLBY, K.M.: Computer Models of Thought and Language, San Francisco, Freeman, 1973.

TAYLOR, R.W. e FRANK, R.L.: CODASYL Database management systems, "ACM Computing Survey", vol. 8, nº 1, 1976

TROST, H. e STEINACKER, I.: The Role of Roles: Some aspects of World Knowledge Representation, Proc. of 7th. IJCAI, Vancouver, 1981.

WINOGRAD, T.: Language as a Cognitive Process, vol. I: Syntax, Addison-Wesley, 1983.

WOODS, W.A: Semantics for a Question Answering Systems, Gorland Pub., 1979.

Informático de AUXINI

Asesor de la Dirección General de

Política Científica